VARIANCE ESTIMATION UNDER REGRESSION IMPUTATION MODEL

Kyuseong Kim, The University of Seoul
Department of Computer Science and Statistics, Seoul, 130-743, Republic of Korea

Key Words: Imputation variance, Prediction, Regression imputation model, Sampling variance, Variance estimation after imputation

1. Introduction

Missing values are common in most sample surveys. Imputation methods are widely used to treat missing values. When a feasible value is inserted in place of missing value by a suitable imputation method, a complete data set is obtained. Then usual statistical analysis programs can be applied directly to the imputed data set and consistent estimator of population mean can be calculated. However, usual variance estimator calculated from imputed data does not cover the imputation variance so that it underestimates true variance of sample mean of imputed data. Therefore, it does lead to incorrect statistical inference. Several authors have given alternative methods or modified methods for consistent variance estimation.

Multiple imputation was proposed by Rubin(1978, 1987) as a way of handling missing data that retains advantages of single imputation and at the same time provides a method of estimating uncertainty due to imputation. Developed from Bayesian point of view, Multiple imputation provides consistent variance estimator so that leads to valid inference from imputed data. However, it needs higher cost of storage and processing than single imputation since it requires multiple complete data set and more seriously, it is available only when imputation method is proper. But proper imputation may not exist in some cases.

An adjusted jackknife variance estimation method was proposed by Rao and Shao(1992) under weighted hot-deck imputation. In this method, jackknife variance estimator is constructed from adjusted jackknife replicates when a respondent is deleted. However, their method is valid only when weighted hot deck imputation method is used. Furthermore, it does not be applied to the multivariate case because of different adjustment for each variable and can not be applied to nonsmooth estimator such as sample quantiles. This method was extended to two phase sampling with ratio and regression imputation and stratified multistage sampling with deterministic imputation. (Rao, 1996; Rao and Sitter, 1995; Sitter, 1997; Sitter and Rao, 1997). For the stratified multistage survey data, Shao, Chen and Chen (1998) proposed an adjusted balanced repeated replication method under imputation. Also Shao and Sitter(1996) showed that correct bootstrap estimates can be obtained by imputing to the bootstrap data set in the same way as the original data set.

A model-based method was suggested by Särndal (1992). He introduced imputation model and decomposed total variance into sampling variance and imputation variance under the model explicitly. Hereafter by estimating two variance components unbiasedly, Särndal obtained unbiased estimator of total variance for the imputed sample mean.

In this paper, we extend the model-based method to more general regression imputation model. In Section 2, we briefly explain basic concept of Särndal's model-based method under imputation model. This method is extended to general regression imputation model in Section 3. Here we suggest a prediction imputation method, which is basically best linear unbiased prediction under the proposed regression imputation model. Furthermore, we propose an unbiased variance estimator of the total variance of sample mean from imputed data. In Section 4, a simulation study is conducted to study the performance of the proposed method compared with unadjusted and adjusted variance estimation methods. Finally, Section 5 gives concluding remarks.

2. Backgrounds

Let A_s be a simple random sample of size n from a finite population, A_r a respondent set of size r among the selected sample and A_{s-r} nonrespondent set of size $n-r$. Denote imputed value for the nonresponse unit k by $\hat{y}_k$. Then an imputed data obtained by using a suitable imputation method is given by

$$y_k^* = \begin{cases} y_k, & k \in A_r \quad \text{: response value} \\ \hat{y}_k, & k \in A_{s-r} \quad \text{: imputed value} \end{cases} \tag{1}$$

From the imputed data, sample mean, $\bar{y}_I$, is

$$\bar{y}_I = \frac{1}{n}\left\{ \sum_{k\in A_r} y_k + \sum_{k\in A_{s-r}} y_k^* \right\} \tag{2}$$

Let p be simple random sampling design, q ignorable response mechanism and ξ imputation model.

Then total error of $\bar{y}_I$ from the population mean $\bar{Y}$ can be decomposed into sampling error, $\bar{y}_s - \bar{Y}$, and imputation error, $\bar{y}_I - \bar{y}_s$. i.e.,

$$\bar{y}_I - \bar{Y} = (\bar{y}_s - \bar{Y}) + (\bar{y}_I - \bar{y}_s)$$

where $\bar{y}_s$ is sample mean from all response.

We say that $\bar{y}_I$ is totally unbiased for the population mean if

$$E_\xi E_p E_q (\bar{y}_I - \bar{Y}) = 0 \tag{3}$$

where each subscript is associated with corresponding probability distribution. In addition, we define total variance of $\bar{y}_I$ as

$$V_{tot} = E_\xi E_p E_q (\bar{y}_I - \bar{Y})^2, \tag{4}$$

which also can be decomposed explicitly into three variance components, sampling variance, V_{sam}, imputation variance, V_{imp}, and covariance, V_{mix}, such as

$$V_{tot} = V_{sam} + V_{imp} + V_{mix} \tag{5}$$

From the above formula, we come to know that if we find unbiased estimators for three variance components, say, v_{sam}, v_{imp} and v_{mix}, then the sum of them

$$v_{tot} = v_{sam} + v_{imp} + v_{mix} \tag{6}$$

become an unbiased estimator of total variance V_{tot}.

3. Prediction Imputation and its Unbiased Variance Estimation

3.1 Regression Imputation Model

We consider a regression imputation model :

$$y_k = x_k'\beta + \varepsilon_k, \ \ \varepsilon_k \ \sim \ (0, v_k\sigma^2), \ k = 1, ..., N \tag{7}$$

where $\beta = (\beta_1, ..., \beta_m)'$ is regression coefficient vervor, $x_k = (x_{k1}, ..., x_{km})'$ is known auxiliary data of unit k, $v_k = v(x_k)$, which is dependent of unit k only, σ^2 is unknown variance imposed on the error ε_k. It is assumed no covariance among errors.

After rearranging response and nonersponse units, we introduce some notations :

- $X_s = (x_{kj}, k \in A_s, j = 1, ..., m)$
- $X_r = (x_{kj}, k \in A_r, j = 1, ..., m)$
- $X_{s-r} = (x_{kj}, k \in A_{s-r}, j = 1, ..., m)$
- $y_r = (y_k, k \in A_r)$: response vector
- $W_r = \text{diag}(v_k, k \in A_r)$: diagonal matrix
- $W_{s-r} = \text{diag}(v_k, k \in A_{s-r})$: diagonal matrix
- $v_{s-r} = (v_k, k \in A_{s-r})$: nonresponse vector

Under the regression imputation model, a missing value may be replaced by the best linear unbiased predicted value as the correct response value. By the prediction theory, we have the best linear unbiased predictor for β from A_r as

$$\hat{\beta}_r = (X_r'W_r^{-1}X_r)^{-1}X_r'W_r^{-1}y_r, \tag{8}$$

so imputed value become $\hat{y}_k = x_k'\hat{\beta}_r$. Thus imputed data set is given by

$$y_k^* = \begin{cases} y_k, & k \in A_r \\ x_k'(X_r'W_r^{-1}X_r)^{-1}X_r'W_r^{-1}y_r, & k \in A_{s-r} \end{cases} \tag{9}$$

Also, sample mean of imputed data is expresed as

$$\bar{y}_I = \frac{1}{n}\{1_r'y_r + 1_{s-r}'X_{s-r}\hat{\beta}_r\} \tag{10}$$

We can show that the imputed sample mean, $\bar{y}_I$, is totally unbiased for the population mean and also find an unbiased estimator of the total variance. The following theorem gives the result without proof.

Theorem 1 *Suppose p is simple random sampling design with sample size n and q is an ignorable response mechanism. Let r be the size of response set. Then under the regression imputation model ξ as in (7), we have the following results :*

(a) *The sample mean from imputed data, $\bar{y}_I$, is totally unbiased for the population mean $\bar{Y}$.*
(b) *An unbiased estimator of total variance of $\bar{y}_I$ is given by*

$$\begin{aligned} v_{tot} \ &= \ \frac{s_I^2}{n} + \frac{1}{n(n-1)}\{(1_s'X_s + 1_r'X_r) \\ &\quad \times(X_r'W_r^{-1}X_r)^{-1}X_{s-r}'1_{s-r} \\ &- \ \sum_{k \in s-r} x_k'(X_r'W_r^{-1}X_r)^{-1}x_k\Bigg\}\hat{\sigma}^2 \end{aligned} \tag{11}$$

where

$$\hat{\sigma}^2 = \frac{1}{r-m}(y_r - X_r\hat{\beta}_r)'W_r^{-1}(y_r - X_r\hat{\beta}_r), \tag{12}$$

s_I^2 is sample variance from the imputed data, and $\hat{\beta}_r$ is given as in (8).

3.2 Ratio Imputation Model

Ratio imputation model is a simple case of regression imputation models, where $m = 1$. We consider

$$y_k = x_k\beta + \varepsilon_k, \quad \varepsilon_k \sim (0, v_k\sigma^2), \; k = 1, ..., N \quad (13)$$

where $x_k > 0$, $v_k = x_k^g$ and $0 \leq g \leq 2$.

Then imputed data is obtained as

$$y_k^* = \begin{cases} y_k, & k \in A_r \\ x_k'(\sum_{k\in r} x_k^{1-g} y_k / \sum_{k \in r} x_k^{2-g}), & k \in A_{s-r} \end{cases} \quad (14)$$

and sample mean from imputed data is given by

$$\bar{y}_I = \frac{1}{n}\left\{\sum_{k\in r} y_k + \sum_{k\in s-r} x_k \left\{\frac{\sum_{k\in r} x_k^{1-g}}{\sum_{k\in r} x_k^{2-g}}\right\}\right\} \quad (15)$$

As a result, unbiased total variance estimator is given by

$$\begin{aligned} v_{tot} \;\; = \;\; & \frac{s_I^2}{n} + \frac{1}{n(n-1)} \\ & \times \left\{ \frac{(\sum_{k\in s} x_k + \sum_{k\in r} x_k) \sum_{k\in s-r} x_k}{\sum_{k\in r} x_k^{2-g}} \right. \\ & \left. - \frac{\sum_{k\in s-r} x_k^2}{\sum_{k\in r} x_k^{2-g}} \right\} \hat{\sigma}^2 \end{aligned} \quad (16)$$

where

$$\hat{\sigma}^2 = \frac{1}{r-1} \sum_{k\in r} \frac{1}{x_k^g} \left\{ y_k - x_k \frac{\sum_{k\in r} x_k^{1-g} y_k}{\sum_{k\in r} x_k^{2-g}} \right\}^2 \quad (17)$$

The usual ratio model is the case of $g = 1$ and the result of $g = 1$ is the same as that of Särndal(1992).

4. A Simulation Study

We conducted a simulation study to examine the performance of the proposed method compared with others. We considered three regression models, simple ratio model (Model 1), simple regression model (Model 2) and multiple regression model (Model 3) with two auxiliary variables. Model 3 is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (18)$$

(β_0=0 and β_2=0 in Model 1 and β_2=0 in Model 2). Auxiliary variable x_1 was generated from gamma distribution $G(g_1, h_1)$ and $x_2|x_1$ was from gamma distribution $G(g_2\ , x_1^{1/5})$, and error ε was from normal distribution $N(0, v\sigma^2)$ independently. For each model, three variance components v were considered. In case of Model 3,

$$v_3 = \left(\frac{1 + x_1 + x_2}{3}\right)^g, \; g = 0, 1.0, 1.5 \quad (19)$$

($v_1 = x_1^g$ in Model 1 and $v_2 = ((1+x_1)/2)^g$ in Model 2)

For each model, we chosen a simple random sample of size $n = 30$ and then we got a respondent set with response rate 70% from the selected sample. Here, we used three response mechanisms, one ignorable($\alpha = 0$) and two nonignorables(α=-0.7, 0.7) in the response probability p :

$$p \propto (\beta_0 + \beta_1 x_1 + \beta_2 x_2)^\alpha \quad (20)$$

As a result, all 27 kinds (3 Models × 3 variance components × 3 response mechanisms) of respondent set were obtained. For each respondent set, we got four imputed data set by employing four imputation methods including mean imputation (M), ratio imputation (R), regression imputation (G) and hot-deck imputation (H).

For comparision study, all 9 kinds of variance estimators were considered. First, four naive variance estimators, v_{IM}, v_{IR}, v_{IG} and v_{IH}, were calculated from imputed data, here subscripts are associated with imputation methods. Next, we considered design-consistent variance estimator (Cochran, 1977) :

$$v_C = \frac{1}{n}s_{ry}^2 + \left(\frac{1}{r} - \frac{1}{n}\right) s_{rd}^2 \quad (21)$$

where $s_{ry}^2 = \sum_{k\in r}(y_k - \bar{y}_r)^2/(r-1)$ and $s_{rd}^2 = \sum_{k\in r}(y_k - (\bar{y}_r/\bar{x}_r)x_k)^2/(r-1)$. It is constructed based on two-phase sampling procedure (first phase - sample selection and second phase - response) and no impūtation is used. The ratio imputation is also used in two-phase sampling procedure, then sample mean after ratio imputation become ratio estimator, i.e.,

$$\bar{y}_I = \frac{1}{r}\sum_{k\in r} y_k + \frac{1}{n-r}\sum_{k\in s-r}\left(\frac{\bar{y}_r}{\bar{x}_r}\right) x_k = \bar{y}_R \quad (22)$$

So, v_C can be an alternative variance estimator under ratio imputation model with $g = 1$. However, if we start with model-based approach under ratio imputation model with $g = 1$, we can expect v_R as in (15) with $g = 1$ to be obtained. Rao and Sitter (1995) considered ratio imputation under two-phase sampling and proposed a jackknife variance estimator, v_{JR}, based on adjusted imputed values, which was shown to have large sample validity under uniform response mechanism.

For an imputed data by hot-deck imputation, Rao and Shao (1992) proposed an unified technique that adjusts jackknife replicates when a respondent is deleted to naive jackknife variance estimator. The adjusted jackknife variance estimator, v_{JH} , is given by

$$v_{JH} = \frac{n-1}{n}\sum_{j=1}^{n}(\bar{y}_I^a(-j) - \bar{y}_I)^2 \qquad (23)$$

where $\bar{y}_I^a$ is adjusted jackknife replicates. Finally, the proposed variance estimator, v_G, is given as in (10) under the assumed regression imputation model.

To evaluate the performance of 9 methods, we generated $B = 100,000$ independent samples for 27 cases. First, we calculated the percent relative bias of the sample mean from imputed data, which is given by

$$REB = \sum_{b=1}^{B}\frac{(\bar{y}_I^{(b)} - \bar{Y})/B}{\bar{Y}} \times 100 \qquad (24)$$

where $\bar{y}_I^{(b)}$ is sample mean from bth replication. Secondly, we got the percent relative bias of the variance estimator,

$$REV = \sum_{b=1}^{B}\frac{(v^{(b)} - V(\bar{y}_I))/B}{V(\bar{y}_I)} \times 100 \qquad (25)$$

where $v^{(b)}$ is variance estimator calculated from the bth replication and $V(\bar{y}_I)$ is estimated variance of $\bar{y}_I$ through simulation. All simulations were performed using SAS/IML, version 6.12. The results are summarized in Table 1, 2 and 3.

It can be noted from Table 1 that under ignorable response mechanisms ($\alpha = 0$), sample means from imputed data have small relative biases. They are not so different. However, under nonignorable response mechanisms, imputed sample means by mean and hot-deck imputation are seriously biased upward (α =-0.7) or downward(α=0.7). But sample means after ratio or regression imputation have less bias than others. In particular, sample mean with regression imputed values is nearly unbiased regardless of ignorable or nonignorable response mechanism.

For the variance estimation, as it can be seen from Table 2 and 3, we can say as a whole that the proposed estimator, v_G, performed better than any other estimators for the most cases and next, two adjusted jackknife variance estimators, v_{JR}, under ratio imputation and v_{JH} under hot-deck imputation were more efficient than other estimators except v_G.

Table 1. Percent Relative Biases of Sample Means from Imputed Data

α	MD	g	$\bar{y}_M$	$\bar{y}_R$	$\bar{y}_G$	$\bar{y}_H$
-0.7	1	0.0	-16.35	-0.01	-0.01	-16.37
		1.0	-16.36	-0.02	-0.02	-16.37
		1.5	-16.37	-0.03	-0.04	-16.37
	2	0.0	-13.14	4.17	0.00	-13.15
		1.0	-13.14	4.18	0.01	-13.14
		1.5	-13.13	4.18	0.01	-13.14
	3	0.0	-9.17	4.34	0.00	-9.20
		1.0	-9.17	4.34	0.01	-9.21
		1.5	-9.17	4.34	0.01	-9.22
0	1	0.0	0.01	-0.01	-0.01	-0.02
		1.0	0.01	-0.01	-0.01	-0.01
		1.5	0.01	-0.01	-0.01	0.00
	2	0.0	0.01	0.15	0.00	0.02
		1.0	0.01	0.16	0.01	0.02
		1.5	0.02	0.16	0.01	0.02
	3	0.0	-0.02	0.37	0.00	-0.02
		1.0	-0.02	0.38	0.00	-0.02
		1.5	-0.01	0.38	0.01	-0.02
0.7	1	0.0	18.95	-0.00	0.00	18.95
		1.0	18.95	0.01	0.01	18.98
		1.5	18.96	0.01	0.01	19.02
	2	0.0	13.20	-2.78	-0.00	13.24
		1.0	13.21	-2.78	-0.00	13.24
		1.5	13.21	-2.78	-0.00	13.24
	3	0.0	9.47	-2.92	0.00	9.46
		1.0	9.47	-2.91	0.01	9.46
		1.5	9.47	-2.92	0.01	9.46

Table 2. Percent Relative Biases of Naive Variance Estimators

α	MD	g	v_{IM}	v_{IR}	v_{IG}	v_{IH}
-0.7	1	0.0	-49.55	-13.68	-9.50	-40.90
		1.0	-50.45	-37.27	-37.27	-41.77
		1.5	-51.05	-50.79	-48.28	-42.36
	2	0.0	-50.65	7.77	-9.22	-42.23
		1.0	-50.86	-12.86	-28.83	-42.52
		1.5	-51.04	-27.90	-41.45	-42.74
	3	0.0	-50.04	5.86	-5.42	-41.98
		1.0	-50.65	-5.33	-19.82	-41.10
		1.5	-50.75	-16.46	-32.02	-41.22
0	1	0.0	-51.78	-8.08	-5.68	-42.88
		1.0	-51.87	-27.03	-27.03	-43.08
		1.5	-51.81	-38.84	-36.68	-43.09
	2	0.0	-51.67	4.93	-7.69	-43.28
		1.0	-51.59	-10.27	-19.75	-43.11
		1.5	-51.55	-22.25	-28.62	-43.01
	3	0.0	-51.58	6.13	-4.87	-43.01
		1.0	-51.50	-3.90	-15.05	-42.96
		1.5	-51.44	-14.03	-23.99	-42.92
0.7	1	0.0	-52.53	-3.37	-2.48	-43.96
		1.0	-52.36	-15.10	-15.10	-43.84
		1.5	-52.12	-23.11	-21.83	-43.64
	2	0.0	-52.32	5.75	-8.69	-43.59
		1.0	-52.21	-4.63	-14.86	-43.39
		1.5	-52.09	-13.34	-17.89	-43.23
	3	0.0	-51.44	10.56	-5.24	-42.67
		1.0	-51.41	1.55	-12.49	-42.59
		1.5	-51.38	-7.62	-17.59	-42.55

Table 3. Percent Relative Biases of Adjusted Variance Estimators

α	MD	g	v_C	v_R	v_{JR}	v_G	v_{JH}
-0.7	1	0.0	-22.20	14.00	0.40	0.16	5.17
		1.0	-26.52	-0.42	-0.87	-0.42	3.60
		1.5	-28.75	-11.37	-1.95	-0.58	2.53
	2	0.0	-25.69	47.79	6.00	0.01	2.79
		1.0	-27.32	29.48	3.83	-0.14	2.29
		1.5	-28.44	14.59	1.95	-0.19	1.90
	3	0.0	-30.50	87.80	4.50	-0.07	3.25
		1.0	-29.35	75.61	4.03	0.35	3.04
		1.5	-28.24	61.04	3.36	0.62	2.83
0	1	0.0	-0.20	7.95	0.09	0.01	1.66
		1.0	-1.10	-0.48	-0.77	-0.48	1.30
		1.5	-1.33	-8.13	-1.27	-0.49	1.27
	2	0.0	0.01	28.52	1.05	0.20	0.89
		1.0	-0.32	17.99	0.55	0.22	1.20
		1.5	-0.56	8.15	-0.03	0.23	1.39
	3	0.0	-3.36	72.97	0.41	-0.20	1.41
		1.0	-2.95	61.98	0.40	0.05	1.49
		1.5	-2.48	48.80	0.30	0.26	1.56
0.7	1	0.0	11.60	2.60	0.25	0.19	-0.28
		1.0	22.84	-0.29	-0.32	-0.29	-0.05
		1.5	30.90	-4.10	-0.62	-0.38	0.29
	2	0.0	15.43	16.23	3.04	-0.14	0.37
		1.0	20.26	10.76	2.07	-0.16	0.73
		1.5	24.31	5.03	1.20	-0.13	1.00
	3	0.0	26.19	59.33	3.40	-0.23	1.99
		1.0	26.36	49.72	2.81	-0.02	2.15
		1.5	26.52	38.32	2.16	0.19	2.21

Specifically, the proposed estimator v_G worked well for most cases of our simulation and will be expected to be highly efficient if the assumed model holds. Furthermore, although our theoretical development of variance and its estimator was derived under ignorable response mechanism only, the simulation results say that the proposed variance estimator can work well under nonignorable response mechanism if the assumed model holds. This is a by-product of the simulation study.

Secondly, as in Table 2, four naive variance estimators underestimated the true variance seriously. it underestimated by over 50 % for most cases in case of mean imputation, and by over 40 % in case of hot deck imputation. Two other estimators, v_{IR} and v_{IG}, also underestimated up to 30 %. They were already expected because the true variance of the estimator from imputed data is greater than that from all observed data due to imputation, but naive variance estimator does not cover the imputation effects by using imputed data as if observed data.

Thirdly, in the first column in Table 3, design consistent variance estimator v_C worked well in case of ignorable response pattern(α=0). By contrast percent relative biases were over 20% in nonignorable cases($\alpha = 0.7, -0.7$). It resulted from the fact that since v_C was constructed under uniform response mechanism, it come to have large variation if the assumption is violated. Fourthly, in the second column of Table 3, v_R, has lower percent bias under the Model 1, i.e., simple ratio model and has large biases under the other models. Because v_R is best only in case of the ratio imputation model with $g = 1$, so it does not works well in the other cases.

Finally, we came to know that two adjusted jackknife variance estimators, v_{JR} and v_{JH} were good candidates for variance estimation from imputed data regardless of response mechanism, ignorable or nonignorable. In particular, the table showed that v_{JR} is very efficient under the first model.

5. Conclusion

In this paper we have studied a prediction method for imputation and its unbiased variance estimation method under regression imputation model. We proposed best linear unbiased predicted value as an imputed value and derived sample mean from imputed data. Also, we obtained unbiased variance estimator of the imputed sample mean under the model. Through a simulation study, we conformed that the sample mean after prediction imputation is nearly unbiased under nonignorable as well as ignorable response mechanism and the proposed variance estimator is more efficient than other known adjusted variance estimators. Furthermore, the proposed estimator is applicable regardless of ignorable or nonignorable response mechanism.

References

Bolfarine, H. and Zacks. S. (1992). *Prediction theory for finite populations.* Springer-Verlag. New York.

Cochran, W.G. (1977). *Sampling techniques.* 3rd ed. Wiley, New York.

Kovar, J. G. (1994). Jackknife variance estimation of imputed survey data. *Survey Methodology*, **20**, 45-52.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, **91**, 499-520.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot -deck imputation. *Biometrika*, **79**, 811-822.

Rao, J. N. K. and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453-460.

Rubin, D.B. (1978). Multiple imputations in sample survey : A phenomenological Bayesian approach to nonresponse. *ASA proceedings of the Section on Survey Research Methods*, 20-28.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley, New York.

Särndal, C.E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18**, 241-252.

Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, **91**, 1278-1288.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, **93**, 819-831.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, **92** , 780-787.

Sitter, R.R. and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *The Canadian Journal of Statistics*, **25** , 61-73.